Sparse reward processes

Christos Dimitrakakis

January 13, 2012

Abstract

We introduce a class of learning problems where the agent is presented with a series of tasks. Intuitively, if there is a relation among those tasks, then the information gained during execution of one task has value for the execution of another task. Consequently, the agent is intrinsically motivated to explore its environment beyond the degree necessary to solve the current task it has at hand. We develop a decision theoretic setting that generalises standard reinforcement learning tasks and captures this intuition. More precisely, we consider a multi-stage stochastic game between a learning agent and an opponent.

We posit that the setting is a good model for the problem of lifelong learning in uncertain environments, where while resources must be spent learning about currently important tasks, there is also the need to allocate effort towards learning about aspects of the world which are not relevant at the moment. This is due to the fact that unpredictable future events may lead to a change of priorities for the decision maker. Thus, in some sense, the model "explains" the necessity of curiosity. Apart from introducing the general formalism, the paper provides algorithms. These are evaluated experimentally in some exemplary domains. In addition, performance bounds are proven for some cases of this problem.

1 Introduction

This paper introduces the setting of sparse reward processes. This captures the essential problem of acting in an unknown environment, with an arbitrary unknown sequence of future objectives. The question is: how do we act so as to achieve the current objective as efficiently as possible, while at the same time acquiring knowledge in order to be able to solve future objectives? Consequently, it analogous to a number of real-world problems. These include both problems in explaining and characterising human and animal behaviour, as well as the design of optimal strategies in problems with high uncertainty about future tasks.

In standard multi-objective problems, one must make decisions that balance the relative importance of objectives. In practice, however, not all of the objectives are applicable all the time and in many cases there may be periods where no objectives are applicable at all. Nevertheless, optimal behaviour may always

be defined with respect to possible *future* objectives. In our framework, the agent assumes that future objectives are unkonwn, and can be *unpredictable*. Thus, while the agent is acting to achieve the current objective, he also acts so as to learn as much about its environment as possible, in order to be able to perform well in any future possible objective.

We formulate this setting in terms of a multi-stage game between a learning agent and an opponent of unknown type. The agent acts within an unknown controlled Markov process, which remains constant (or more generally, is drawn from the same distribution) at every stage. In addition, at each stage of the game, a payoff function is chosen by the opponent, which determines the agent's utility. Loosely speaking, the agent must act not only so as to maximise expected utility at each stage, but also so that he can be better prepared for whatever payoff function the opponent will select at the next stage. We call such problems *sparse reward processes*.

Our first technical contribution is a measure-theoretic formulation of the payoff action that defines each stage. This allows us to relax the usual Markovian assumptions with respect to rewards without necessarily making planning intractable. In addition, we show that, when the opponent is nature, the environment can be described as an unknown MDP. Finally, we show that, when the opponent is adversarial, a nearly optimal strategy is to maximise the information gain with respect to the MDP model, linking our formulation to exploration heuristics such as compression progress and approximations to the value of information [10, sec. 23.7].

Multi-armed bandit problems with covariates [17, 21, 13, 15] is a closely related setting, where again the payoff function is given at the beginning of every stage. In that setting, however, the opponent is always nature and, more importantly, the only thing observed after an action is chosen is a noisy reward signal. So, in some sense, it is a harder problem than the one considered herein (and indeed [15] prove a lower bound). Consequently, the main difference between this setting and the covariate bandits (as well as the related multi-task bandit [12] setting) is that there is always an underlying, observable, environment which the agent can explore and learn about.

2 Setting

At its simplest, the setting can be formalised as a multi-stage game between the agent, an opponent, and nature. At the beginning of the k-th stage the opponent chooses a payoff ρ_k, which he reveals to the agent, who then selects a policy π_k. The agent's expected utility for that stage is $V_k \triangleq V(\rho_k, \pi_k)$. If the environmental dynamics are known for all stages, then it is straightforward, but not necessarily trivial, to act so as to maximise the expected payoff across all stages, by playing the optimal strategy for each stage and disregarding the remaining stages. When the environment dynamics are *unknown, but related* in some way across stages, then learning about the environment is important for performing well in the later stages. The setting then becomes an interesting

special case of the exploration-exploitation problem.

The remainder of this section presents the setting in more detail. Section 2.1 defines the environment that the agent is acting in. Section 2.2 introduces the payoff function that the opponent chooses before each stage. Section 2.3 discusses the policy and the resulting value of the game between the agent and the opponent. Finally, Section 2.4 puts all elements together in the formulation of *sparse reward processes*.

2.1 The environment

As mentioned in the introduction, at every stage, the agent is acting within an unknown environment, in order to maximise the expectation of a known payoff function. The payoff function is chosen by an opponent, who however has no control over the environment's dynamics. For simplicity, we make the assumption that the environment's dynamics are constant throughout all stages.[1] More specifically, we define the environment to be a controlled Markov process:

Definition 1. *A controlled Markov process (CMP) $\nu \in \mathcal{N}$ is a tuple $\nu = (\mathcal{S}, \mathcal{A}, \mathcal{T})$, with state space $\mathcal{S}$, action space $\mathcal{A}$, and transition kernel*

$$\mathcal{T} \triangleq \{ \tau(\cdot \mid s, a) \mid s \in \mathcal{S}, a \in \mathcal{A} \},$$

indexed in $\mathcal{S} \times \mathcal{A}$ such that $\tau(\cdot \mid s, a)$ is a probability measure[2] on $\mathcal{S}$. The CMP ν defines a discrete-time Markov process: If at time t the environment is in state $s_t \in \mathcal{S}$ and the agent chooses action $a_t \in \mathcal{A}$, then the next state s_{t+1} is drawn with a probability independent of previous states and actions:

$$\mathbb{P}_\nu(s_{t+1} \in S \mid s^t, a^t) = \tau(S \mid s_t, a_t) \qquad S \subset \mathcal{S}. \tag{2.1}$$

In the above, and throughout the text, we use the following conventions. We employ $\mathbb{P}_\nu$ to denote the probability of events under a process ν, while we use $s^t \equiv s_1, \ldots, s_t$ and $a^t \equiv a_1, \ldots, a_t$ to represent sequences of variables. Similarly $\mathcal{S}^t$ denotes product spaces, and $\mathcal{S}^* \triangleq \bigcup_{t=0}^{\infty} \mathcal{S}^t$ denotes the set of all sequences of states.

Throughout this paper, we assume that the transition kernel (and possibly the state and action spaces) is not known to the agent, who must estimate it through interaction. On the other hand, the payoff function, chosen by the opponent, is revealed to the agent at the beginning of each stage.

2.2 The payoff

At the k-th stage of the game, a payoff function ρ_k chosen by the opponent. The payoff function simply encodes how desirable a state sequence is to the agent for the particular task. In particular if $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^*$ are two state sequences, then

[1] However, this assumption can be relaxed.

[2] We assume the measurability of all sets with respect to some appropriate σ-algebra. This will usually be the Borel algebra $\mathcal{B}(X)$ of the set X.

$\mathbf{s}$ is preferred to $\mathbf{s}'$ in round k if and only if $\rho_k(\mathbf{s}) \geq \rho_k(\mathbf{s}')$. While in this paper we assume that the opponent has knowledge of what the payoffs ρ_k are, we also later discuss how to relax this to an agnostic opponnet.

The payoff functions are somewhat more general than the usual reinforcement learning (RL) settings. Recall that in reinforcement learning the agent is acting within a Markov decision process μ (MDP). This is essentially a CMP equipped with a set of distributions $\{ R(\cdot \mid s) \mid s \in \mathcal{S} \}$ on rewards $r_t \in \mathbb{R}$, such that $\mathbb{P}_\mu(r_t \in B \mid s_t = s) = R(B \mid s)$ for any B in the Borel sets of $\mathbb{R}$. In the *infinite-horizon, discounted reward* setting, the utility is defined as the discounted sum of rewards $\sum_t \gamma^t r_t$, where $\gamma \in [0, 1]$ is a discount factor. We can map this to our framework, by setting:

$$\rho(s^T) = \sum_{t=1}^{T} \gamma^t \, \mathbb{E}(r_t \mid s_t) = \sum_{t=1}^{T} \gamma^t \int_{-\infty}^{\infty} r \, \mathrm{d}R(r \mid s_t) \tag{2.2}$$

to be the utility of a sequence of states s^T. Note that in our case, making the payoff function stochastic does not usefully generalise our problem, since it is revealed to the agent at the beginning of each stage.

2.3 The policy

The payoff ρ_k is revealed to the decision maker, who then chooses a policy π_k, which he uses to interact with the environment. The controlled Markov process and the payoff function jointly define a Markov decision process [14] (MDP), denoted by $\mu_k = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_k)$. The agent's policy π_k selects actions with distribution $\pi_k(a_t \mid s^t)$, meaning that the policy is not necessarily stationary. Together with the Markov decision process μ_k, it defines a distribution on the sequence of states, such that:

$$\mathbb{P}_{\mu_k, \pi_k}(s_{t+1} \in S \mid s^t) = \int_{\mathcal{A}} \tau(S \mid a, s_t) \, \mathrm{d}\pi(a \mid s^t). \tag{2.3}$$

This interaction results in a (random) sequence of states $\mathbf{s}$, whose utility U_k to the agent is:

$$U_k \triangleq \rho_k(\mathbf{s}), \qquad \mathbf{s} \in \mathcal{S}^*. \tag{2.4}$$

Since we consider that the payoff function is explicitly revealed to the agent, making payoffs stochastic does not usefully generalise our setting. However, there is still randomness due to the fact that we sequence of states is random. We set the value of each stage to the *expected utility*:

$$V_k \triangleq V(\rho_k, \pi_k) \triangleq \mathbb{E}_{\nu, \pi_k} U_k \tag{2.5}$$

The agent tries to maximise $\sum_k V_k$, the total expected utility across stages.

2.4 Sparse reward processes

The complete sparse reward process is a special case of a stochastic game.[19] However, we are particularly interested in processes where only few state sequences have payoffs. We model this by mapping each payoff function to a *finite* measure on $\mathcal{S}^*$. A simple way to capture this intuition formally is the following:

Definition 2. *A sparse reward process is a multi-stage stochastic game with K stages, where the k-th stage is a Markov decision process $\mu_k = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_k)$, whose payoff function $\rho_k : \mathcal{S}^* \to [0, 1]$, is revealed to the agent not later than after $k-1$ stage is complete. The agent chooses policy π_k, with expected utility $V_k \triangleq V(\rho_k, \pi_k)$. The Markov decision process terminates at time t and the stage ends, with fixed termination probability q.*

The process is called sparse *if there exists a measure σ on $\mathcal{S}^*$ such that, for every $\rho_k \in \mathcal{R}$, the payoff measure λ_k on $\mathcal{S}^*$, defined as:*

$$\lambda_k(S) = \int_S \rho_k(\mathbf{s}) \, \mathrm{d}\sigma(\mathbf{s}), \qquad \forall S \subset \mathcal{S}^*, \tag{2.6}$$

satiisfies $\lambda_k(\mathcal{S}^*) \leq 1$. The agent's goal is to find a sequence π_k maximising $\sum_{k=1}^{K} V_k$.

Some discussion of this definition is in order. Firstly, the game is structured such that the opponent does not necessarily have complete knowledge of the underlying Markov process (which is out of his control) or the payoffs (which he selects). Secondly, the agent's choice of policy depends on assumptions about the opponent. Thirdly, the fixed termination probability q is equivalent to an infinite horizon discounted reward reinforcement learning problem [see 14]. Finally, the last condition ensures that the opponent cannot place arbitrarily large rewards in certain parts of the space, and so cannot make the task arbitrarily difficult.

In fact, the payoff measure construction not only results in a natural measure of the sparseness of payoffs, but it also enables much of the subsequent technical development, through the following rather trivial, but important lemma:

Lemma 1. *Given a payoff function ρ for which there exists a payoff measure λ satisfying the conditions of Def. 2 for some σ, the utility of any policy π on the MDP $\mu = (\nu, \rho)$, can be written as:*

$$\mathbb{E}_{\pi,\mu} U = \int_{\mathcal{S}^*} p_{\pi,\nu}(\mathbf{s}) \, \mathrm{d}\lambda(\mathbf{s}), \tag{2.7}$$

where $p_{\pi,\nu}$ is the probability (density) of $\mathbf{s}$ (with respect to σ) under the policy π and the environment ν.

Proof.

$$\mathbb{E} U = \int_{\mathcal{S}^*} \rho(\mathbf{s}) \, \mathrm{d}P_{\pi,\nu}(\mathbf{s}) = \int_{\mathcal{S}^*} \rho(\mathbf{s}) p_{\pi,\nu}(\mathbf{s}) \, \mathrm{d}\sigma(\mathbf{s}) = \int_{\mathcal{S}^*} p_{\pi,\nu}(\mathbf{s}) \, \mathrm{d}\lambda(\mathbf{s}) \tag{2.8}$$

□

3 Optimality conditions

The optimality of a given policy for the agent depends on the assumptions made regarding the opponent. In a worst-case setting, it is natural to view each stage as a zero-sum game between us and the opponent, where the agent's gain is the opponent's loss. If the opponent is nature, then the sparse reward process can be seen as an MDP. This is also true in the case where we employ a prior over the opponent's type.

We begin by introducing some additional concepts. Firstly, let us define the oracle policy for stage k:

Definition 3 (Oracle stage policy)**.** *Given the process ν and the payoff ρ_k at stage k, the optimal policy π^*_{ν,ρ_k} is:*

$$\pi^*_{\nu,\rho_k} \triangleq \arg\max_{\pi} \int_{\mathcal{S}^*} \rho_k(\mathbf{s}) \, \mathrm{d}P_{\pi,\nu}(\mathbf{s}). \tag{3.1}$$

This policy is normally unattainable by the agent, since ν is unknown. Instead, we assume that the agent maintains a probabilistic belief ξ over the set of CMPs $\mathcal{N}$, such that $\xi(B) = \int_B \mathrm{d}\xi(\nu)$ is the belief that $\nu \in B$, for $B \subset \mathcal{N}$. In a purely Bayesian setting, such a belief is subjective, as it is based on an arbitrary prior ψ. However, if it is actually known that the CMP was drawn from some distribution ψ, then the posterior belief:

$$\xi(B \mid D) = \frac{\int_B \prod_{t,i} P_\nu(\mathbf{s}_{i,t+1} \mid \mathbf{s}_{i,t}, \mathbf{a}_{i,t}) \, \mathrm{d}\psi(\nu)}{\int_{\mathcal{N}} \prod_{t,i} P_\nu(\mathbf{s}_{i,t+1} \mid \mathbf{s}_{i,t}, \mathbf{a}_{i,t}) \, \mathrm{d}\psi(\nu)} \tag{3.2}$$

conditioned on a set of state-action observation sequences $D = \{ \mathbf{s}_i, \mathbf{a}_i \}$ from the process, has good estimation properties.[6, 16, 11, 1, 4]

3.1 When the opponent is nature

Consider the case when the opponent selects the payoffs ρ_k by drawing them from some fixed, but unknown distribution with measure $\phi(\cdot \mid \theta)$, parametrised by $\theta \in \Theta$, such that:

$$\mathbb{P}(\rho_k \in B) = \phi(B \mid \theta), \qquad \forall k \in \{ 1, 2, \ldots, K \}, \ \forall B \subset \mathcal{R}. \tag{3.3}$$

In that case, the Bayes-optimal strategy for the agent is to maintain a belief ω on the joint space of CMPs and $\Theta \times \mathcal{N}$ and solve the problem with backwards induction [3], if possible. This is because of the following fact:

Theorem 1. *When the opponent is Nature, the SRP is an MDP.*

Proof. We prove this by construction. For a set of reward functions $\mathcal{R}$, the state space of the MDP can be factored into the reward function and the state of the dynamics, so $\mathcal{S} = \mathcal{R} \times \mathcal{S}_0$. If there are K reward functions, we can write the state space as $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}_k$. Let the action space be $\mathcal{A}$, such that there is a

one-to-one mapping $M_{ij} : \mathcal{S}_i \leftrightarrow \mathcal{S}_k$. In addition, for any i, j all states $s \in \mathcal{S}_i$, the transition probabilities obey:

$$\mathbb{P}(s_{t+1} \in B \mid s_t = s, a_t = a) = \mathbb{P}(s_{t+1} \in B \mid s_t = M_{i,j}(s), a_t = a) \tag{3.4}$$

$$\mathbb{P}(s_{t+1} \in \mathcal{S}_j \mid s_t \in \mathcal{S}_i, a_t = a) = \begin{cases} q, & j \neq i, \\ (1-q), & j = i, \end{cases} \tag{3.5}$$

for any $B \subset \mathcal{S}$. It is easy to verify that this is in agreement with Def. 2. □

Unfortunately, in most cases, the Bayes-optimal solution is intractable [8, 3, 6].

3.2 When the opponent is adversarial

The expected utility of any policy π given a belief ξ over $\mathcal{N}$, is:

$$\mathbb{E}_{\xi,\pi} U = \int_{\mathcal{N}} \left(\int_{\mathcal{S}^*} U(\mathbf{s}) \, \mathrm{d}P_{\pi,\nu}(\mathbf{s}) \right) \mathrm{d}\xi(\nu). \tag{3.6}$$

Let P^*_ν and P^*_ξ be the probability measures on $\mathcal{S}^*$ arising from the optimal policy given the full CMP ν and given a particular belief ξ over CMPs respectively, assuming known payoffs ρ. The opponent can take advantage of our partial knowledge and select a payoff function that maximises our loss relative to the optimal policy:

$$\ell_k(\xi, \mu) \triangleq \max_\lambda \int_{\mathcal{S}^*} (P^*_\nu - P^*_\xi) \, \mathrm{d}\lambda_k. \tag{3.7}$$

The above definition clearly implies that the opponent should reduce payoffs in sets of state sequences which have a much higher probability under ν compared to under ξ. To make this non-trivial for the opponent, we have restricted the payoff functions to $\rho(\mathcal{S}^*) \leq 1$.

Theorem 2. *Consider a* two-stage *game, where there is no reward received during the first stage, i.e. $\rho_1(\mathbf{s}) = 0$ for all $\mathbf{s}$. Then, it is sufficient to choose the policy maximising the expected information, in order to mininise our expected regret.*

Proof. See that:

$$\ell_k(\xi, \nu) \leq \max_\lambda \int_{\mathcal{S}^*} |P^*_\nu - P^*_\xi| \, \mathrm{d}\lambda_k \leq \|P^*_\nu - P^*_\xi\|_\sigma. \tag{3.8}$$

Thus, choosing a policy that maximises the expected information gain, minimises the expected worst-case loss at the next stage. □

This natural result is in broad agreement with past ideas of relating curiosity to gaining knowledge about the environment (e.g. the work starting

with [18]). Consequently, pure information-gathering strategies can have good quality guarantees in this two-stage adversarial game.

For more general games, we must employ other strategies, however, as we need to balance information gathering (exploration) with obtaining rewards in the current stage (exploitation). Unfortunately, even for the two-stage game, finding the policy that maximises the expected information gain is in the same complexity class as finding the Bayes-optimal policy. For this reason, in the next section we consider upper confidence bound algorithms, in the spirit of UCB [2] and UCRL [9], which, although approximate, perform quite well.

3.3 Algorithms

Here we present two simple algorithms for SRPs. The first, Upper Confidence bound SRP (UCSRP, Alg. 1) chooses policies based on simple confidence bounds, similarly to UCB. The second, Bayesian Thompson sampling (SRP, Alg. 2), chooses a policy by drawing samples from a posterior distribution.

In order to simplify the exposition, we restrict our attention to some arbitrary stage k and consider a setting where we have a finite set of policies $\mathcal{P}$. Each policy $\pi \in \mathcal{P}$, coupled with the unknown dynamics, defines a probability measure $P_\pi(S) \triangleq \mathbb{P}_{\pi,\nu}(\mathbf{s} \in S)$. Let D be a metric between probability measures on $\mathcal{S}^*$, i.e.:

$$D(P, Q) = \int_{\mathcal{S}^*} |P(\mathbf{s}) - Q(\mathbf{s})| \, \mathrm{d}\sigma(\mathbf{s}). \tag{3.9}$$

For any policy π, let the corresponding empirical measure on $\mathcal{S}^*$ be $\hat{P}_\pi$, and let:

$$\mathcal{Q}_\epsilon(\hat{P}_\pi) \triangleq \left\{ Q \;\middle|\; D(Q, \hat{P}_\pi) \leq \epsilon \right\},$$

be a confidence region around the empirical measure. Then we define:

$$P_\pi^+ \triangleq \operatorname*{arg\,max}_{Q : \in \mathcal{Q}_\epsilon(\hat{P}_\pi)} \mathbb{E}_Q \, \rho \tag{3.10}$$

to be the measure within the interval maximising the expected payoff. For any i, we can define a signed measure c_π: $P_\pi^+ = \hat{P}_\pi + c_\pi$. In addition, there always exists an optimal choice π^*, such that $\mathbb{E}_{P_{\pi^*}} \rho \geq \mathbb{E}_{P_\pi} \rho$ for all π.

Algorithm 1 UCSRP: Upper Confidence bound SRP

1: Select the smallest $\{\mathcal{Q}_\epsilon(\hat{P}_\pi) \mid \pi \in \mathcal{P}\}$ s.t. $\mathbb{P}(\exists \pi : P_\pi \notin \mathcal{Q}_\epsilon(\hat{P}_\pi) \leq 1/k)$,
2: Choose π_k such that $\mathbb{E}(\rho_k \mid \pi_k) \geq \mathbb{E}(\rho_k \mid \pi)$ for all π.
3: Execute π_k, observe outcome $\mathbf{s}$ and payoff $\rho(\mathbf{s})$.
4: Update $\{\hat{P}_\pi \mid \pi \in \mathcal{P}\}$.

Algorithm 2 BTSRP: Bayesian Thompson sampling SRP

1: Set initial beliefs $\xi_0(P_\pi)$.
2: **for** $k = 1, \dots, K$ **do**
3: For all π, sample $\hat{P}_\pi \sim \xi_k(P_\pi)$.
4: Choose π_k such that $\mathbb{E}_{\hat{P}_\pi} \rho_k \geq \mathbb{E}_{\hat{P}_j} \rho_k$ for all j.
5: Execute π_k, observe outcome $\mathbf{s}$ and payoff $\rho(\mathbf{s})$.
6: Calculate $\xi_{k+1}(\cdot) \triangleq \xi_k(\cdot \mid \pi, \mathbf{s})$.
7: **end for**

Lemma 2. *Consider a payoff function ρ with corresponding payoff measure λ. Assume that ϵ is such that confidence regions hold, i.e. that $P_\pi \in \mathcal{Q}_\epsilon(\hat{P}_\pi)$ for all i. For UCSRP to choose a sub-optimal policy π, it sufficient that:*

$$\mathbb{E}(\rho \mid P_{\pi^*}) \leq \mathbb{E}(\rho \mid P_\pi) + 2 \int c_\pi \, \mathrm{d}\lambda.$$

Proof. Since UCSRP always chooses π maximising P^+_π, if we choose a sub-optimal π then it must hold that $\mathbb{E}(\rho \mid P^+_{\pi^*}) \leq \mathbb{E}(\rho \mid P^+_\pi)$. Since the confidence regions hold, $\mathbb{E}(\rho \mid P_{\pi^*}) \leq \mathbb{E}(\rho \mid P^+_{\pi^*})$, $\mathbb{E}(\rho \mid P_{\pi^*}) \leq \mathbb{E}(\rho \mid P^+_{\pi^*})$ and $\mathbb{E}(\rho \mid \hat{P}_\pi) \leq \mathbb{E}(\rho \mid P_\pi + c_\pi)$. Consequently:

$$\mathbb{E}(\rho \mid P_{\pi^*}) \leq \mathbb{E}(\rho \mid P^+_\pi) = \int (\hat{P}_\pi + c_\pi) \, \mathrm{d}\lambda \leq \int (P_\pi + c_\pi) \, \mathrm{d}\lambda = \mathbb{E}(\rho \mid P_\pi) + 2 \int c_\pi \, \mathrm{d}\lambda$$

□

Theorem 3. *Let $c_{\pi,k}$ be the relevant signed measure for policy π in stage k. Assume that $\|c_{\pi,k}\|^b \leq a n_{\pi,k}$, with $n_{\pi_k} = \sum_{i=1}^k \mathbb{I}\{\pi_i = \pi\}$.*

Proof. Let $\pi^*_k \triangleq \arg\max_{\pi \in \mathcal{P}} \rho_k' \pi$ be the optimal policy at stage k in hindsight, and let π_k be our policy for that stage. Then the regret after K stages, $\mathcal{L}_K$, is bounded as follows:

$$\begin{aligned}
\mathcal{L}_K &\leq \max_\rho \sum_{k=1}^K \rho_k' \pi^*_k - \rho_k' \pi_k = \max_\rho \sum_{k=1}^K \rho_k' \pi^*_k - \rho_k' \sum_{\pi \in \mathcal{P}} \mathbb{I}\{\pi_k = \pi\} \\
&\leq \sum_{\pi \in \mathcal{P}} \sum_{k=1}^K \mathbb{I}\{\pi_k = \pi\} \max_{\rho_k} \rho_k'(\pi^*_k - \pi) \\
&\leq \sum_{\pi \in \mathcal{P}} \sum_{k=1}^K \max_{\rho_k} \mathbb{I}\left\{ 2 \int c_{\pi,k} \geq \rho_k'(\pi^*_k - \pi_k) \right\} \rho_k'(\pi^*_k - \pi).
\end{aligned}$$

□

The actual shape of the confidence region, for UCSRP, and the belief, for BTSRP, depend on the model we are using. In general, they have the form $c_i = a n_i^{1/b}$, where n_i is the number of times the i-th policy was chosen and $a > 0$, and $b \geq 1$, but can be tighter if there is an interrelationship between policies. In this paper, we consider two types of games.

3.3.1 Associative stages

There are a total of K stages. In each stages, the agent starts from an initial state s_0, then selects a policy π_k, which takes him to some state s with probability $\mathbb{P}_{\nu,\pi_k}(s|s_0)$, which is unknown. Then the agent receives a reward $\rho(s)$ and the stage immediately terminates. The state space is discrete, and there are $N < \infty$ policies. So, in a sense, this problem is a variant of the contextual bandit problem, or the problem of prediction with side-information. In this case, the context, or side-information is the reward function. There are, however, fundamental technical differences between this problem and the standard contextual bandit setting.

For this game, if we assume independent policies, confidence intervals for UCSRP can be constructed via Weissman's bound on the L1 norm of deviations of emprical estimates of multinomial distributions [20]. Then, for each arm i, with probability at least $1 - \delta$, the true transition probability is within the L1 ball of radius:

$$c_i = \sqrt{2[(|\mathcal{S}| - 1) \ln 2 - \ln \delta]/n_i}, \tag{3.11}$$

around our empirical estimate. Similarly, for the BTSRP policy, we maintain a product-Dirichlet distribution (see for example [3]) on the outcomes of actions.

3.3.2 Markov stages

Again, there are a total of K stages. In each stage, the agent starts from a uniformly drawn state $s \in \mathcal{S}$, then selects a policy π_k. The environment is Markov, and the stage terminates with fixed probability q, known to the agent. Then the opponent selects a payoff for the next stage.

Once more, this game is similar to bandits with side-information. However, now the different arms (policies) are no longer independent. For this game, we can again employ the Weissman bound or a Dirichlet distribution for each state-action pair if all sets are finite.

3.4 Experiments

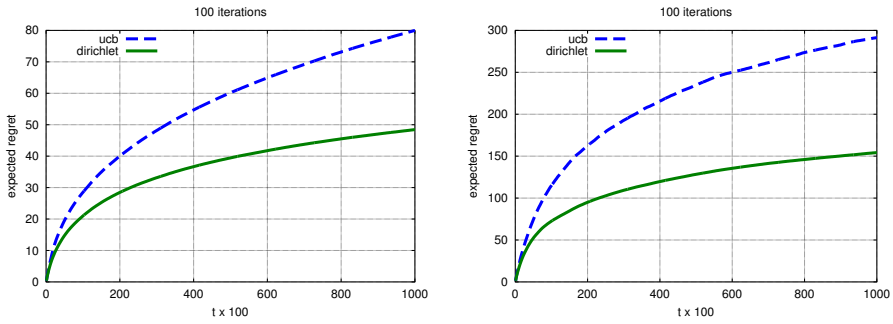
We compared UCSRP with BTSRP, as well as the greedy policy for both games.

There are two variants of the games. In the first, the opponent is *nature*, i.e. reward functions are sampled uniformly on the simplex $\{\rho \in R | \sum_s \rho(s) = 1\}$ for every stage.

In the second variant, the opponent is *adversarial*. This opponent has knowledge of P_i, and also maintains the empirical distributions $\hat{P}_i$. The ssumption is that the agent, whatever its method, will use something close to the empirical distributions anyway. Then the reward function maximising

$$\max_i \mathbb{E}_{P_i} \rho - \mathbb{E}_{P_j} \rho, \qquad j = \arg\max_{i=1,\dots,N} \hat{P}_\pi \rho$$

is chosen.



(a) Nature (b) Adversarial

Figure 1: Comparison of the expected cumulative regret after k stages, between UCSRP and BTSRP, on a problem with a multi task bandit problem, for two types of opponents. The first type, Fig. 1(a), is nature. There, the reward at the next stage is drawn from a fixed distribution (the $\mathcal{D}\mathit{irichlet}(1)$ distribution on the reward simplex). The second type, Fig. 1(b), is adversarial. There, the reward is chosen so that maximal expected regret will be incurred at the current stage.

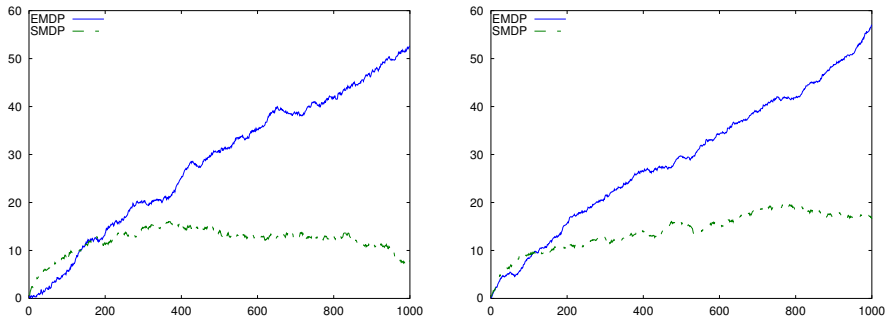
For discrete spaces, there is always a finite number of deterministic policies, each corresponding to a unique distribution on the states, and consequently on the payoffs. After a stage terminates, we move to the next stage. Consequently, the overall problem is very similar to a linear context bandit problem, where ρ_k is the side-information.

The results for the *associative stages* problem are shown in Fig. 1. It seems that the UCSRP performance is quite good, with the BTSRP algorithm being even better, with a Dirichlet prior. In both cases, performance is severely degraded when the opponent is adversarial (Fig. 1(b)) compared to when it is nature (Fig. 1(a)).

For the *Markov stages* setting, results are shown in Fig. 2, for an adversarial opponent and a stopping probability of $q = 0.5$ at every timestep of each stage. There, we compare BTSRP with the stationary greedy policy, i.e. the stationary policy which maximises payoff for the current stage in empirical expectation. As can be seen clearly in both cases, the regret suffered by the greedy policy grows linearly, while that of BSRP grows negligibly.

4 Conclusion and related work

We introduced the setting of *sparse reward processes*, which captures the essential problem of acting in an unknown environment with arbitrarily selected future objective. As such, it is a good surrogate for a number of real-world problems. These include both problems in explaining and characterising human



(a) 2 states (b) 4 states

Figure 2: Comparison of the expected cumulative regret after k stages, between BTSRP and stagewise stationary greedy policies, on randomly generated MDPs, with 2 and 4 states respectively, against an adversarial opponent. The greedy policy suffers linear cumulative regret.

and animal behaviour, as well as the design of optimal strategies in problems with high uncertainty about future tasks.

We have shown that, in certain special cases of the game, a good strategy is to maximise the expected information gain. This links the problem to previous work on curiosity. In addition, we have shown how the problem is, in a special case, a standard Markov decision process. Finally, we have evaluated two simple algorithms and shown that they perform well on this problem.

Naturally, similar ideas have appeared in the literature previously. The most closely related setting are multi-armed bandit problems with covariates [17, 21, 13]. Specifically, [17] considers a one-armed bandit problem in a Bayesian setting, for an exponential family mode, and proves that a myopic policy is asymptotically optimal, in a discounted setting. Yang and Zhu [21] uses a non-parametric regression model for estimation and an ϵ-greedy policy. The main assumptions are that the payoff functions are drawn from a known distribution. There is also the recent work of [15], which prove a lower bound on the regret.

Finally, there are a relation to multi-task learning, such as [12], and in particular to learning with multiple bandits [5, 7], which consider the problem of finding the optimal policies for a number of sub-problems. The starting states of Dimitrakakis and Lagoudakis [5] and the bandit problems of Gabillon et al. [7] are loosely analogous to the different payoff functions seen here.

Acknowledgements

Many thanks to Peter Auer, for his insights on reinforcement learning and intrinsic rewards, Constantin Rothkopf, discussions with whom spurred this particular line of investigation. Finally, I would like to thank Boi Faltings, for comments

on the draft paper, as well as Andrew Barto, Jürgen Schmidhuber and Jochen Triesch for their perspectives on multi-task learning, curiosity and compression progress and to Thomas Léauté for proofreading.

A Auxiliary results

Lemma 3. *If σ is a measure on (X, Σ) and $\rho : X \to \mathbb{R}_{*+}$ is a Σ-measurable function, then*

$$\lambda(A) = \int_A \rho(x)\,\mathrm{d}\sigma(x)$$

is a measure.

Proof. Let $A, B \in \Sigma$ with $A \cap B = \emptyset$.

$$\lambda(A \cup B) = \int_{A\cup B} \rho(x)\,\mathrm{d}\sigma(x) = \int_A \rho(x)\,\mathrm{d}\sigma(x) + \int_B \rho(x)\,\mathrm{d}\sigma(x) = \lambda(A) + \lambda(B).$$

The non-negativity of λ follows easily from the fact that $\rho(x) \geq 0\ \forall x \in X$.

$$\lambda(A) = \int_A \rho(x)\,\mathrm{d}\sigma(x) \geq \inf_{x\in A} \rho(x)\sigma(A) \geq 0.$$

□

References

[1] J. Asmuth, L. Li, M. L. Littman, A. Nouri, and D. Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *UAI 2009*, 2009.

[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002.

[3] Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970.

[4] Christos Dimitrakakis. Complexity of stochastic branch and bound methods for belief tree search in Bayesian reinforcement learning. In *2nd international conference on agents and artificial intelligence (ICAART 2010)*, pages 259–264, Valencia, Spain, 2009. ISNTICC, Springer.

[5] Christos Dimitrakakis and Michail G. Lagoudakis. Rollout sampling approximate policy iteration. *Machine Learning*, 72(3):157–171, September 2008. doi: 10.1007/s10994-008-5069-3. Presented at ECML'08.

[6] Michael O'Gordon Duff. *Optimal Learning Computational Procedures for Bayes-adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts at Amherst, 2002.

[7] Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. In *NIPS 2011*, 2011.

[8] C. J. Gittins. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, New Jersey, US, 1989.

[9] Thomas Jacksh, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[10] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[11] J. Zico Kolter and Andrew Y. Ng. Near-Bayesian exploration in polynomial time. In *ICML 2009*, 2009.

[12] Gábor Lugosi, Omiros Papaspiliopoulos, and Gilles Stoltz. Online multi-task learning with hard constraints. In *COLT 2008*, 2008.

[13] N.G. Pavlidis, D.K. Tasoulis, and D.J. Hand. Simulation studies of multi-armed bandits with covariates. In *Tenth International Conference on Computer Modeling and Simulation*, pages 493–498. IEEE, 2008.

[14] Marting L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New Jersey, US, 2005.

[15] P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT*, pages 54–66. Omnipress, 2010.

[16] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive POMDPs. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, Cambridge, MA, 2008. MIT Press.

[17] Jyotirmoy Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, 19(4):pp. 1978–2002, 1991. ISSN 00905364. URL `http://www.jstor.org/stable/2241915`.

[18] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. 1991.

[19] L. S. Shapley. Stochastic games. *PNAS*, pages 1095–1100, 1953.

[20] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M.J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[21] Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):pp. 100–121, 2002. ISSN 00905364. URL `http://www.jstor.org/stable/2700004`.